



SEGMENTASI PELANGGAN PADA PLATFORM MARKETPLACE MENGUNAKAN K-MEANS DAN DBSCAN BERBASIS PERILAKU TRANSAKSI RFM

Sanusi Amir

Ucilinkun9@gmail.com

Article Info

Article history:

Received: 03-02-2025

Revised: 20-03-2025

Accepted: 10-04-2025

Published: 15-04-2025

Keywords:

segmentasi pelanggan,
RFM, K-Means,
DBSCAN, marketplace.

ABSTRACT

Segmentasi pelanggan berbasis *Recency-Frequency-Monetary* (RFM) umumnya dilakukan dengan K-Means, tetapi pada marketplace dengan pembeli sekali transaksi yang sangat dominan, frekuensi pembelian menjadi terdegenerasi dan pelanggan ekstrem dapat mendistorsi centroid. Penelitian ini membandingkan K-Means dan DBSCAN serta mengusulkan pendekatan hibrida DBSCAN→K-Means pada 93.357 pelanggan unik dari 96.477 pesanan terkirim pada marketplace Olist (September 2016–Agustus 2018). Fitur RFM ditransformasi logaritmik (F dan M) dan distandardisasi. Jumlah kluster K-Means ditentukan dengan Elbow, Silhouette, Davies-Bouldin Index (DBI), dan Calinski-Harabasz Index (CHI), sedangkan parameter DBSCAN ditentukan dari titik lutut grafik *k-distance* dan pencarian *grid*. Evaluasi mencakup validitas internal, stabilitas (*Adjusted Rand Index*/ARI pada 20 subsampel 80%), keseimbangan segmen, dan kontribusi pendapatan. K-Means dengan $k = 4$ dipilih (CHI tertinggi; Silhouette tertinggi untuk $k \geq 3$). DBSCAN ($\epsilon = 0,264$; $\text{minPts} = 20$) menghasilkan satu kluster dominan berisi 96,7% pelanggan sehingga tidak diskriminatif, meskipun DBI-nya rendah (0,576). Pendekatan hibrida mengisolasi 509 pelanggan ekstrem (0,55%) dan membentuk lima segmen dengan validitas setara K-Means (Silhouette 0,375; DBI 0,770), namun lebih stabil (ARI rerata 0,994 vs 0,988; $p = 0,0096$) dan menghindari kontaminasi centroid: rerata monetary kluster pelanggan berulang pada K-Means terlalu tinggi 20,2%. Segmen pembeli tunggal bernilai tinggi (29,7% pelanggan) menyumbang 56,0% pendapatan sehingga menjadi prioritas konversi ke pembelian kedua. DBSCAN lebih tepat berperan sebagai penyaring pelanggan ekstrem daripada pengganti K-Means.

Kata kunci: segmentasi pelanggan; RFM; K-Means; DBSCAN; marketplace

1. INTRODUCTION

Pertumbuhan platform *marketplace* menghasilkan data transaksi dalam jumlah besar yang merekam perilaku pembelian setiap pelanggan. Bagi pengelola platform, data tersebut menjadi dasar untuk membedakan perlakuan pemasaran: pelanggan yang bernilai tinggi perlu dipertahankan, pelanggan yang mulai tidak aktif perlu diaktifkan kembali, dan pelanggan bernilai rendah perlu dilayani dengan biaya yang efisien. Model *Recency-Frequency-Monetary* (RFM) banyak dipakai untuk tujuan ini karena hanya membutuhkan tiga atribut yang tersedia pada hampir semua sistem transaksi, yaitu kebaruan, keseringan, dan nilai belanja [1]-[2]-[3].

Sebagian besar penelitian segmentasi RFM menggunakan K-Means. Fauzan dan Alfian [4] melaporkan bahwa K-Means mengungguli K-Medoids dan *Fuzzy C-Means* pada data *e-commerce* publik dan menghasilkan tiga segmen (loyal, perlu perhatian, dan menjanjikan). Alzami dkk. [5] dan

Tabianan dkk. [6] juga menerapkan K-Means pada data pembelian *e-commerce*, sedangkan Samidi dkk. [7] menunjukkan bahwa nilai aktual RFM menghasilkan kluster yang lebih baik daripada skor RFM. John dkk. [8] membandingkan beberapa algoritma klusterisasi pada data ritel daring. Namun, Yulisasih dkk. [9] mengingatkan bahwa K-Means sering menghasilkan kluster yang luas dan kurang rinci sehingga evaluasi validitas kluster tidak dapat diabaikan.

Algoritma berbasis kepadatan, terutama DBSCAN, menawarkan kemampuan yang tidak dimiliki K-Means, yaitu menemukan kluster berbentuk sembarang dan menandai titik *noise*. Irawan dkk. [10] menerapkan DBSCAN pada RFM data *e-commerce* untuk mendeteksi pelanggan pencilan dan tidak aktif. Perbandingan langsung antara kedua paradigma memberi hasil yang beragam. Saputra dan Yusuf [11] menemukan K-Means lebih unggul pada Silhouette dan DBI untuk pengguna Transjakarta,

tetapi DBSCAN mampu mengidentifikasi pencilan dengan nilai moneter tinggi. Agustino dkk. [12] melaporkan bahwa *Affinity Propagation* lebih baik daripada DBSCAN pada tenant inkubator bisnis. Ullah dkk. [13] memperoleh dua klaster dari DBSCAN dengan Silhouette sedikit lebih tinggi, tetapi memilih tiga klaster karena dua kelompok dinilai kurang mendalam.

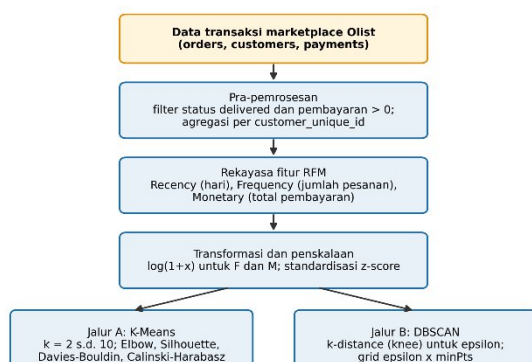
Dari telaah tersebut terdapat beberapa celah. Pertama, struktur data marketplace yang didominasi pembeli sekali transaksi, sehingga atribut frekuensi hampir biner, jarang dibahas secara eksplisit. Kedua, DBSCAN dan K-Means umumnya diposisikan sebagai pesaing, padahal DBSCAN dapat berperan komplementer sebagai penyaring pelanggan ekstrem sebelum K-Means. Ketiga, evaluasi banyak bertumpu pada indeks validitas internal, sementara stabilitas segmen, yang misalnya dinilai dengan ARI oleh Ho dkk. [14], dan konsentrasi pendapatan jarang dilaporkan bersamaan. Keempat, pemilihan parameter DBSCAN sering tidak didokumentasikan secara sistematis.

Penelitian ini bertujuan (1) membandingkan K-Means dan DBSCAN pada segmentasi RFM marketplace berskala 93.357 pelanggan, (2) mengusulkan pendekatan hibrida DBSCAN→K-Means dengan parameter DBSCAN yang ditentukan dari titik lutut grafik *k-distance*, dan (3) mengevaluasi ketiga pendekatan dengan protokol multikriteria yang mencakup validitas internal, stabilitas, keseimbangan segmen, dan kontribusi pendapatan. Kebaruan penelitian terletak pada (i) rancangan hibrida yang mengubah *noise* DBSCAN menjadi segmen pelanggan ekstrem yang dapat ditindaklanjuti, (ii) diagnosis efek frekuensi terdegenerasi dan artefak indeks validitas pada DBSCAN, serta (iii) protokol evaluasi yang mengaitkan kualitas klaster dengan konsekuensi bisnis. Pernyataan kebaruan ini terbatas pada literatur yang ditelaah dalam naskah ini.

2. MATERIALS AND METHODS

2.1 Kerangka Penelitian

Penelitian ini berjenis kuantitatif-eksperimental dan mengikuti tahapan pada Gambar 1: pengumpulan data, pra-pemrosesan, rekayasa fitur RFM, transformasi dan penskalaan, pemodelan tiga jalur (K-Means, DBSCAN, dan hibrida), evaluasi, serta *profiling* segmen.



Gambar 1. Kerangka penelitian

2.2 Dataset

Data yang digunakan adalah Brazilian E-Commerce Public Dataset by Olist, yaitu data transaksi marketplace yang tersedia secara publik dan telah dianonimkan.

(<https://www.kaggle.com/datasets/olistbr/brazilian-e-commerce>). Tiga tabel dipakai: pesanan (*orders*), pelanggan (*customers*), dan pembayaran (*payments*). Dataset memuat 99.441 pesanan. Pesanan dipertahankan apabila berstatus *delivered* (96.478 pesanan) dan memiliki nilai pembayaran lebih dari nol, sehingga tersisa 96.477 pesanan pada periode 2016-10-03 sampai 2018-08-29. Karena *customer_id* pada dataset bersifat per pesanan, identitas pelanggan diambil dari *customer_unique_id*, sehingga diperoleh 93.357 pelanggan unik. Nilai pembayaran dinyatakan dalam Real Brasil (BRL).

2.3 Rekayasa Fitur dan Transformasi

Tanggal acuan ditetapkan sehari setelah transaksi terakhir (2018-08-30). Untuk setiap pelanggan, *Recency* (R) adalah selisih hari antara tanggal acuan dan transaksi terakhir, *Frequency* (F) adalah jumlah pesanan berbeda, dan *Monetary* (M) adalah total pembayaran. Mengikuti temuan Samidi dkk. [7] bahwa nilai aktual RFM lebih baik daripada skor kuantil, penelitian ini memakai nilai aktual, bukan skor 1–5. Karena F dan M sangat menjulur ke kanan (skewness 11,09 dan 9,21), keduanya ditransformasi dengan logaritma, sedangkan R tidak ditransformasi karena skewness-nya rendah (0,45) dan transformasi logaritma justru memperburuknya (-1,25). Seluruh fitur kemudian distandardisasi.

$$x' = \ln(1 + x) \quad (1)$$

$$z = (x' - \mu) / \sigma \quad (2)$$

2.4 K-Means

K-Means dijalankan dengan jarak Euclidean, inisialisasi *k-means++*, dan 10 inisialisasi acak (seed 42) untuk *k* = 2 sampai 10. Jumlah klaster dipilih dengan empat kriteria: metode Elbow (inersia), koefisien Silhouette, DBI, dan CHI. Silhouette

(persamaan 3) dihitung pada sampel acak 10.000 pelanggan karena kompleksitas komputasinya kuadratik terhadap jumlah data (Shutaywi & Kachouie, [15] sedangkan DBI dan CHI dihitung pada seluruh data.

$$s(i) = (b(i) - a(i)) / \max\{a(i), b(i)\} \quad (3)$$

dengan $a(i)$ rerata jarak titik ke anggota klasternya dan $b(i)$ rerata jarak ke klaster terdekat lain.

2.5 DBSCAN

DBSCAN mengelompokkan titik berdasarkan kepadatan menggunakan dua parameter, yaitu radius ϵ dan jumlah minimum tetangga minPts. Titik yang tidak terjangkau dari titik inti mana pun diberi label *noise*. Untuk menghindari pemilihan parameter yang sewenang-wenang (lihat penekanan dokumentasi parameter pada Irawan dkk [10], nilai ϵ ditentukan dari grafik *k-distance*: jarak setiap titik ke tetangga ke-minPts diurutkan menaik, dan titik lutut ditetapkan pada titik yang berjarak terjauh dari garis lurus yang menghubungkan titik awal dan akhir kurva ternormalisasi. Kandidat minPts adalah 6, 10, 20, 50, dan 100. Sebagai uji kepekaan, dilakukan pencarian *grid* dengan $\epsilon = 0,10; 0,15; 0,20; 0,25; \text{ dan } 0,30$. Konfigurasi akhir memakai minPts = 20 dengan ϵ pada titik lutut (0,264).

2.6 Pendekatan Hibrida DBSCAN→K-Means

Pendekatan hibrida yang diusulkan terdiri atas empat langkah:

1. Jalankan DBSCAN pada fitur RFM terstandarisasi dengan parameter dari Bagian 2.5.
2. Tetapkan seluruh titik *noise* sebagai satu segmen *pelanggan ekstrem* yang ditangani secara terpisah.
3. Jalankan K-Means ($k = 4$, sama dengan K-Means baku agar perbandingan terkendali) hanya pada titik inti.
4. Gabungkan hasil menjadi 5 segmen dan lakukan *profiling*.

2.7 Protokol Evaluasi

Evaluasi mencakup empat aspek. (a) Validitas internal berupa Silhouette, DBI, dan CHI. Agar adil, seluruh metode dinilai pada titik inti (bukan *noise*). (b) Stabilitas: data disubsampel 80% sebanyak 20 kali, setiap metode dijalankan ulang, dan label dibandingkan dengan solusi acuan pada titik yang sama menggunakan ARI, seperti praktik penilaian stabilitas pada Ho dkk.[14]. Untuk DBSCAN, minPts diskalakan 0,8 kali. Perbedaan ARI antara hibrida dan K-Means diuji dengan uji peringkat bertanda Wilcoxon satu arah berpasangan. (c) Keseimbangan segmen berupa proporsi klaster terbesar dan entropi Shannon ternormalisasi. (d) Kontribusi pendapatan setiap segmen terhadap total pembayaran.

2.8 Lingkungan Komputasi

Analisis dilakukan dengan Python 3.12, scikit-learn 1.8.0, pandas 3.0.2, dan SciPy 1.17.1. Seluruh proses acak memakai seed tetap sehingga hasil dapat direproduksi; kode analisis disertakan sebagai berkas pendukung.

3. RESULTS AND DISCUSSION

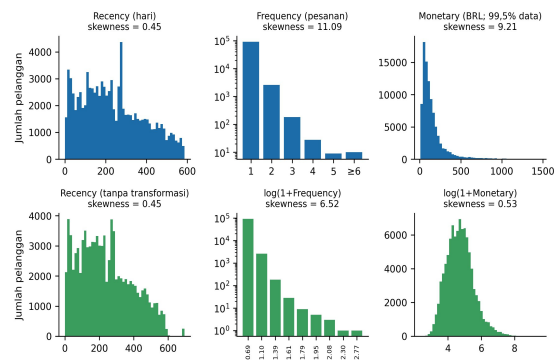
3.1 Karakteristik Data dan Struktur RFM

Tabel 1. Statistik deskriptif fitur RFM ($n = 93.357$)

Fitur	Min.	Median	Rerata	Maks.	Skewness (asli)	Skewness (transformasi)
Recency (hari)	0	218	237,5	695	0,45	0,45*
Frequency (pesanan)	1	1	1,033	15	11,09	6,52
Monetary (BRL)	9,59	107,78	165,20	13.664,08	9,21	0,53

*Recency tidak ditransformasi.

Tabel 1 dan Gambar 2 menunjukkan struktur data yang khas. Sebanyak 97,0% pelanggan (90.556 orang) hanya bertransaksi satu kali dan hanya 2.801 pelanggan yang bertransaksi dua kali atau lebih. Rerata monetary (BRL 165,20) jauh di atas mediannya (BRL 107,78), yang menandakan ekor kanan yang panjang. Konsekuensinya, atribut frekuensi hampir biner: hampir seluruh variasi F terjadi antara nilai 1 dan nilai ≥ 2 . Kondisi ini berbeda dengan penelitian pada data yang didominasi pelanggan berulang, dan berimplikasi langsung pada pemilihan k dan perilaku DBSCAN. Temuan ini sejalan dengan Pangestu dkk.[2], yang juga mendapati klaster berloyalitas rendah didominasi pelanggan satu kali beli.

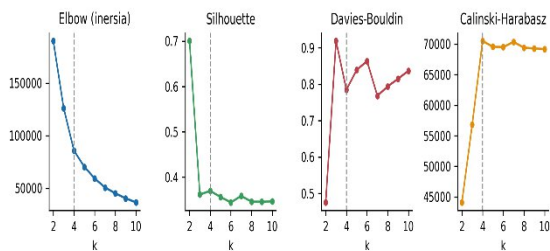


Gambar 2. Distribusi R, F, dan M sebelum (atas) dan sesudah (bawah) transformasi

3.2 Hasil K-Means dan Penentuan Jumlah Klaster

Tabel 2. Hasil evaluasi K-Means untuk berbagai nilai k

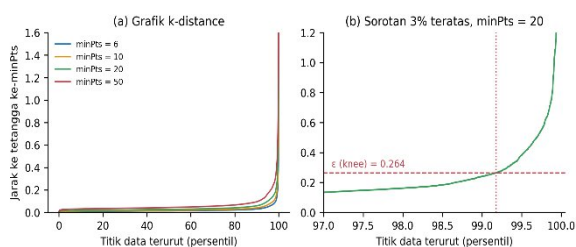
k	Inersia	Silhouette	DBI	CHI
2	190.194	0,701	0,476	44.115
3	126.270	0,362	0,919	56.855
4	85.768	0,370	0,785	70.497
5	70.372	0,356	0,839	69.545
6	59.317	0,345	0,864	69.485
7	50.722	0,359	0,768	70.351
8	45.145	0,346	0,794	69.396
9	40.376	0,346	0,815	69.272
10	36.526	0,347	0,837	69.159



Gambar 3. Kurva Elbow, Silhouette, DBI, dan CHI untuk $k = 2$ sampai 10 (garis putus-putus: $k = 4$)

Silhouette tertinggi (0,701) dan DBI terendah (0,476) terjadi pada $k = 2$. Namun profil kluster menunjukkan bahwa solusi ini hanya memisahkan pelanggan $F = 1$ (90.556 pelanggan) dari pelanggan $F \geq 2$ (2.801 pelanggan), sehingga tidak memberikan informasi segmentasi yang dapat ditindaklanjuti. Hal serupa dicatat Agustino dkk. [12] saat mengulas penelitian terdahulu yang hanya menghasilkan dua segmen dengan frekuensi sebagai satu-satunya pembeda. Untuk $k \geq 3$, CHI mencapai puncak pada $k = 4$ (70.497) dan Silhouette tertinggi juga pada $k = 4$ (0,370), dengan DBI 0,785. Karena itu $k = 4$ dipilih. Kasus $k = 2$ menunjukkan bahwa pemilihan k semata-mata berdasarkan satu indeks dapat menyesatkan, yang sejalan dengan penekanan Yuliasih dkk. [9] pada evaluasi kluster serta penggunaan banyak indeks oleh Ullah dkk. [13].

3.3 Hasil DBSCAN



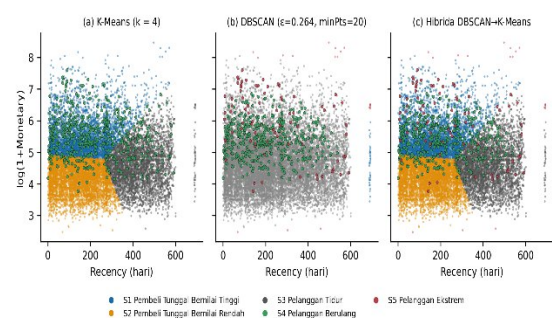
Gambar 4. Grafik k -distance untuk beberapa nilai minPts (a) dan sorotan 3% teratas pada $\text{minPts} = 20$ dengan titik lutut (b)

Tabel 3. Sebagian hasil pencarian grid DBSCAN (baris tebal: konfigurasi terpilih)

minPts	ϵ	Kluster	Noise (%)	Kluster terbesar (%)	Silhouette	DBI
6	0,200	15	0,31	96,7	0,421	0,543
10	0,200	6	0,52	96,7	0,408	0,495
20	0,100	20	3,72	95,2	0,238	0,534
20	0,150	8	1,54	96,3	0,309	0,602
20	0,200	7	0,87	96,5	0,347	0,518
20	0,264	3	0,55	96,7	0,417	0,576
20	0,300	3	0,47	96,7	0,417	0,578
50	0,200	3	2,47	96,1	0,678	0,569
100	0,250	2	3,32	95,9	0,694	0,324
100	0,300	2	2,15	96,3	0,692	0,360

Titik lutut grafik k -distance untuk $\text{minPts} = 6, 10, 20, 50$, dan 100 berturut-turut berada pada $\epsilon = 0,139; 0,168; 0,264; 0,361; 0,440$ (Gambar 4). Pada konfigurasi terpilih ($\text{minPts} = 20; \epsilon = 0,264$), DBSCAN menghasilkan 3 kluster dan 509 titik *noise* (0,55%). Kluster terbesar memuat 90.236 pelanggan (96,7%), kluster kedua 2.383 pelanggan ($F = 2,00$), dan kluster ketiga hanya 229 pelanggan dengan Recency rerata 691 hari. Kluster terakhir ini merupakan pelanggan paling awal pada data dan terbentuk karena adanya celah aktivitas transaksi pada akhir 2016, sehingga lebih mencerminkan artefak temporal daripada segmen perilaku.

Pada seluruh kombinasi parameter yang diuji, kluster terbesar selalu memuat sekitar 87% sampai 97% pelanggan (Tabel 3). Nilai Silhouette (0,414) dan DBI (0,576) DBSCAN tampak lebih baik daripada K-Means, tetapi keduanya dihitung tanpa titik *noise* dan didominasi satu kluster besar, sehingga tidak mencerminkan kegunaan segmentasi. Temuan ini konsisten dengan Ullah dkk. [13] yang menolak solusi DBSCAN dua kluster meskipun Silhouette-nya sedikit lebih tinggi, dan dengan Saputra dan Yusuf [11] yang menyimpulkan bahwa DBSCAN lebih berguna untuk mengidentifikasi pencilan bernilai tinggi.



Gambar 5. Visualisasi hasil klusterisasi pada bidang Recency dan $\log(1+\text{Monetary})$ (sampel acak 15.000 pelanggan): K-Means (a), DBSCAN (b), dan hibrida (c). Titik bertepi hitam adalah pelanggan dengan $F \geq 2$. Warna (a) dan (c) mengikuti legenda; pada (b) abu-abu = kluster dominan, hijau = kluster $F = 2$, biru = kluster Recency ± 690 hari, merah = noise

3.4 Pelanggan Ekstrem dan Pendekatan Hibrida

Pelanggan yang ditandai *noise* oleh DBSCAN berjumlah 509 orang (0,55%) dengan rerata monetary BRL 964,43, atau 5,8 kali rerata seluruh pelanggan, dan hanya menyumbang 3,2% pendapatan. Kelompok ini heterogen: 44,8% bertransaksi tiga kali atau lebih, 28,7% berbelanja lebih dari BRL 1.000, dan 17,9% hanya bertransaksi sekali tetapi bernilai tidak lazim. Heterogenitas ini alasan mereka lebih tepat ditangani secara individual, misalnya melalui manajemen akun

atau verifikasi kemungkinan penjual ulang, daripada digabung ke segmen massal.

Pada data inti, K-Means dengan $k = 4$ menghasilkan Silhouette 0,377, DBI 0,770, dan CHI 68.348. CHI pada data inti naik landai untuk $k > 4$ (69.328 pada $k = 5$ hingga 75.440 pada $k = 8$), tetapi Silhouette tertinggi untuk $k \geq 3$ tetap pada $k = 4$. Karena itu dan demi perbandingan terkendali, $k = 4$ dipertahankan.

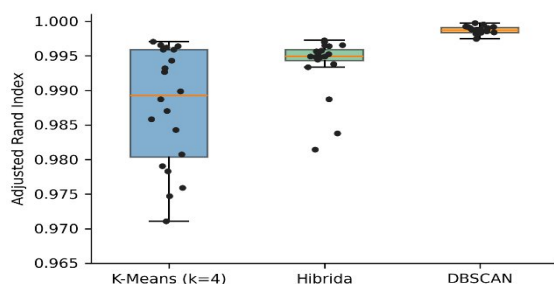
3.5 Perbandingan Metode

Tabel 4. Perbandingan K-Means, DBSCAN, dan hibrida

Kriteria	K-Means (k = 4)	DBSCAN	Hibrida
Jumlah segmen (termasuk noise)	4	4	5
Silhouette (data inti)	0,375	0,414	0,375
DBI (data inti)	0,769	0,576	0,770
CHI (data inti)	68.348	17.811	68.348
Segmen terbesar (% pelanggan)	38,2	96,7	38,2
Entropi ternormalisasi	0,860	0,122	0,751
ARI stabilitas (rerata $\pm$ SD)	0,988 $\pm$ 0,008	0,999 $\pm$ 0,001	0,994 $\pm$ 0,004
ARI stabilitas (minimum)	0,971	0,997	0,981

Validitas internal dihitung pada titik inti (bukan noise) agar sebanding. Silhouette memakai sampel 10.000 titik.

Tabel 4 menunjukkan tiga hal. Pertama, validitas internal hibrida praktis identik dengan K-Means (ARI antara kedua solusi pada titik inti 0,987), sehingga pendekatan hibrida tidak boleh diklaim memperbaiki indeks validitas. Kedua, DBSCAN tampak unggul pada DBI dan Silhouette, tetapi proporsi kluster terbesarnya 96,7% dan entropinya hanya 0,122, sehingga tidak diskriminatif. Skor ARI DBSCAN yang sangat tinggi (0,999) pun sebagian besar mencerminkan dominasi satu kluster. Ketiga, hibrida lebih stabil daripada K-Means baku: ARI rerata 0,994 berbanding 0,988, dengan simpangan baku yang kira-kira setengahnya dan minimum yang lebih tinggi (0,981 vs 0,971). Uji Wilcoxon satu arah berpasangan signifikan ($W = 167$; $p = 0,0096$), dan hibrida lebih stabil pada 15 dari 20 ulangan (Gambar 6). Selain itu, *noise* DBSCAN cukup konsisten antarsubsampel (indeks Jaccard rerata 0,937).



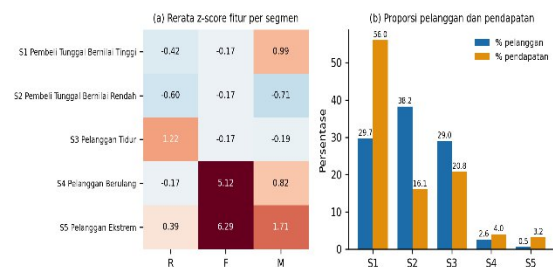
Gambar 6. Sebaran ARI pada 20 subsampel 80% untuk ketiga pendekatan

Manfaat praktis utama pendekatan hibrida adalah mencegah kontaminasi centroid (Tabel 5). Pada K-Means baku, 418 dari 509 pelanggan ekstrem (82,1%) masuk ke kluster pelanggan berulang. Akibatnya rerata monetary kluster tersebut terlalu tinggi 20,2% dibandingkan segmen pelanggan berulang pada hibrida, dan rerata frekuensinya juga terinflasi. Bagi pengelola platform, ini berarti anggaran program loyalitas yang dihitung dari profil K-Means akan melebihi-lebihkan nilai pelanggan berulang yang sesungguhnya.

Tabel 5. Kontaminasi centroid kluster pelanggan berulang

Ukuran	K-Means (k = 4)	Hibrida (S4)	Selisih
Jumlah pelanggan	2.801	2.383	418
Rerata Frequency	2,11	2,00	+5,7%
Rerata Monetary (BRL)	308,59	256,64	+20,2%
Pelanggan ekstrem yang tercampur	418	0	-

3.6 Profil Segmen dan Implikasi Manajerial



Gambar 7. Profil segmen hibrida: rerata z-score fitur setelah transformasi (a) serta proporsi pelanggan dan pendapatan (b)

Tabel 6. Profil lima segmen hasil pendekatan hibrida

Segmen	Pelanggan	% Pelanggan	R (hari)	F	M (BRL)	% Pendapatan
S1 Pembeli Tunggal Bernilai Tinggi	27.701	29,67	174	1,00	311,98	56,04
S2 Pembeli Tunggal Bernilai Rendah	35.697	38,24	146	1,00	69,39	16,06
S3 Pelanggan Tidur	27.067	28,99	424	1,00	118,25	20,75
S4 Pelanggan Berulang	2.383	2,55	211	2,00	256,64	3,97
S5 Pelanggan Ekstrem	509	0,55	296	2,45	964,43	3,18

Profil pada Tabel 6 memperlihatkan bahwa nilai bisnis terkonsentrasi pada pembeli tunggal. Segmen S1 (pembeli tunggal bernilai tinggi) hanya 29,7% dari pelanggan tetapi menyumbang 56,0% pendapatan, dengan Recency rerata 174 hari. Mengubah sebagian dari mereka menjadi pembeli kedua adalah tuas terbesar. Segmen S3 (pelanggan tidur; Recency rerata 424 hari, lebih dari satu tahun) mencakup 29,0% pelanggan dengan kontribusi 20,8%; reaktivasinya diperkirakan lebih sulit (tanpa data respons kampanye, hal ini masih berupa dugaan) sehingga perlu dihitung

biaya-manfaatnya. Segmen S2 (bernilai rendah) merupakan kelompok terbesar (38,2%) tetapi hanya menyumbang 16,1%, sehingga cocok untuk intervensi otomatis berbiaya rendah. Pelanggan berulang (S4) sangat sedikit (2,6%), yang mengindikasikan bahwa retensi merupakan tantangan utama platform. Rekomendasi strategi dirangkum pada Tabel 7 dan selaras dengan pola retargeting dan reaktivasi yang dibahas Irawan dkk. [10].

Tabel 7. Rekomendasi strategi pemasaran per segmen

Segmen	Karakteristik	Rekomendasi strategi
S1 Pembeli Tunggal Bernilai Tinggi	Belanja besar, satu kali, kontribusi pendapatan terbesar	Kampanye konversi ke pembelian kedua (kupon terbatas waktu, rekomendasi produk terkait); pantau kepuasan pascapembelian.
S2 Pembeli Tunggal Bernilai Rendah	Belanja kecil, satu kali, jumlah terbesar	Bundling dan ambang gratis ongkir untuk menaikkan nilai keranjang; komunikasi otomatis berbiaya rendah.
S3 Pelanggan Tidur	Tidak aktif lebih dari satu tahun	Reaktivasi selektif dengan insentif terukur; hentikan komunikasi jika tidak ada respons untuk menekan biaya.
S4 Pelanggan Berulang	Pernah membeli ulang, jumlah kecil	Program loyalitas dan keanggotaan; pertahankan dengan perhatian prioritas.
S5 Pelanggan Ekstrem	Sangat heterogen, terpisah dari pola massal	Tinjauan individual: manajemen akun untuk yang bernilai tinggi, verifikasi kemungkinan penjual ulang atau anomali.

3.7 Keterbatasan

Penelitian ini memiliki beberapa keterbatasan. Pertama, data Olist berasal dari marketplace Brasil dan mencakup September 2016 sampai Agustus 2018, sehingga bukan data terbaru dan tidak mewakili perilaku konsumen Indonesia; replikasi pada data marketplace Indonesia yang lebih mutakhir diperlukan sebelum temuan digeneralisasi. Kedua, hanya tiga fitur RFM yang dipakai, tanpa kategori produk atau demografi seperti pada model RFM yang diperluas (Ho dkk., 2023). Ketiga, Silhouette dihitung pada sampel 10.000 titik. Keempat, ukuran segmen ekstrem peka terhadap parameter DBSCAN: pada minPts = 20, proporsi *noise* berkisar 0,47% ($\epsilon = 0,30$) hingga 3,72% ($\epsilon = 0,10$). Kelima, evaluasi bersifat internal tanpa label acuan dan tanpa uji lapangan atas efektivitas strategi.

4. CONCLUSION

Pada 93.357 pelanggan marketplace yang 97,0%-nya membeli sekali, K-Means dengan $k = 4$ menghasilkan segmen yang layak dipakai, sedangkan DBSCAN membentuk satu klaster dominan (96,7%) sehingga tidak diskriminatif meskipun indeks validitasnya tampak baik. Pendekatan hibrida DBSCAN→K-Means mempertahankan validitas internal K-Means, meningkatkan stabilitas secara signifikan (ARI 0,994 vs 0,988; $p = 0,0096$), dan menghindari inflasi centroid sebesar 20,2% pada klaster pelanggan berulang dengan mengisolasi 509 pelanggan ekstrem. Lima segmen yang terbentuk menunjukkan bahwa 56,0% pendapatan berasal dari 29,7% pelanggan pembeli tunggal bernilai tinggi, sehingga konversi ke pembelian kedua menjadi prioritas strategis. Penelitian selanjutnya disarankan mereplikasi pendekatan ini pada data marketplace Indonesia yang lebih baru, menambah fitur (kategori produk, metode pembayaran, ulasan), mencoba HDBSCAN untuk mengurangi kepekaan parameter,

serta menguji efektivitas strategi segmen melalui eksperimen A/B.

ACKNOWLEDGEMENTS

Penulis mengucapkan terima kasih kepada ITB STIKOM Ambon atas dukungan fasilitas penelitian, serta kepada Olist dan André Sionek yang telah menyediakan dataset publik pada platform Kaggle.

REFERENCES

- [1] Alamsyah, P. E. Prasetyo, Sunyoto, S. H. Bintari, D. D. Saputro, S. Rohman, and R. N. Pratama, "Customer segmentation using the integration of the recency frequency monetary model and the K-Means cluster algorithm," *Scientific Journal of Informatics*, vol. 9, no. 2, pp. 189–196, 2022, doi: 10.15294/sji.v9i2.39437.
- [2] P. I. Pangestu, T. I. Hermanto, and D. Irmayanti, "Analisis segmentasi pelanggan berbasis Recency Frequency Monetary (RFM) menggunakan algoritma K-Means," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 3, pp.1486–1492,2023,doi: 0.36040/jati.v7i3.7171.
- [3] D. S. Nugraha, I. Thoib, N. Sururi, F. Bayu F, and B. P. Candra, "Segmentasi pelanggan berbasis analisis RFM menggunakan algoritma K-Means clustering," *Jurnal Komputer Teknologi Informasi Sistem Komputer (JUKTISI)*, vol. 4, no. 2, pp. 1361–1369, 2025, doi: 10.62712/juktisi.v4i2.641.
- [4] R. M. Fauzan and G. Alfian, "Segmentasi pelanggan e-commerce menggunakan fitur Recency, Frequency, Monetary (RFM) dan algoritma klasterisasi K-Means," *JISKA (Jurnal Informatika Sunan Kalijaga)*, vol. 9, no. 3, pp. 170–177,2024,doi: 0.14421/jiska.2024.9.3.170-177.
- [5] F. Alzami, F. D. Sambasri, M. Nabila, R. A. Megantara, A. Akrom, R. A. Pramunendar, D. P. Prabowo, and P. Sulistiyawati, "Implementation of RFM method and K-Means algorithm for customer segmentation in e-commerce with Streamlit," *ILKOM Jurnal Ilmiah*, vol. 15, no. 1, pp.32–44,2023,doi: 10.33096/ilkom.v15i1.1524.32-44.
- [6] K. Tabianan, S. Velu, and V. Ravi, "K-Means clustering approach for intelligent customer segmentation using customer purchase behavior data," *Sustainability*, vol. 14, no. 12, p. 7243, 2022, doi: 10.3390/su14127243.
- [7] Samidi, R. Y. Suladi, and D. Kusumaningsih, "Comparison of the RFM model's actual value and score value for clustering," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 7, no. 6, pp. 1430–1438, 2023, doi: 10.29207/resti.v7i6.5416.
- [8] J. M. John, O. Shobayo, and B. Ogunleye, "An exploration of clustering algorithms for customer

- segmentation in the UK retail market," *Analytics*, vol. 2, no. 4, pp. 809–823, 2023, doi: 10.3390/analytics2040042.
- [9] B. N. Yuliasih, H. Herman, S. Sunardi, and H. Yuliansyah, "Evaluation of K-Means clustering using silhouette score method on customer segmentation," *ILKOM Jurnal Ilmiah*, vol. 16, no. 3, pp. 330–342, 2024, doi: 10.33096/ilkom.v16i3.2325.330-342.
- [10] F. Irawan, F. Amaluddin, and A. Wijayanti, "E-commerce customer segmentation based on RFM analysis using DBSCAN algorithm to improve marketing strategy," *Jurnal EMT KITA*, vol. 9, no. 4, pp. 1850–1859, 2025, doi: 10.35870/emt.v9i4.5375.
- [11] Saputra and R. Yusuf, "Perbandingan algoritma DBSCAN dan K-Means dalam segmentasi pelanggan pengguna transportasi publik Transjakarta menggunakan metode RFM," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 4, pp. 1346–1361, 2024, doi: 10.57152/malcom.v4i4.1516.
- [12] D. P. Agustino, I. G. B. A. Budaya, I. G. Harsemadi, I. K. Dharmendra, and I. M. S. Astika Pande, "Perbandingan algoritma DBSCAN dan Affinity Propagation dalam segmentasi pelanggan tenant inkubator bisnis," *Jurnal SISFOKOM (Sistem Informasi dan Komputer)*, vol. 12, no. 2, pp. 315–321, 2023, doi: 10.32736/sisfokom.v12i2.1682.
- [13] Ullah, M. I. Mohmand, H. Hussain, S. Johar, I. Khan, S. Ahmad, H. A. Mahmoud, and S. Huda, "Customer analysis using machine learning-based classification algorithms for effective segmentation using Recency, Frequency, Monetary, and Time," *Sensors*, vol. 23, no. 6, p. 3180, 2023, doi: 10.3390/s23063180.
- [14] T. Ho, S. Nguyen, H. Nguyen, N. Nguyen, D.-S. Man, and T.-G. Le, "An extended RFM model for customer behaviour and demographic analysis in retail industry," *Business Systems Research*, vol. 14, no. 1, pp. 26–53, 2023, doi: 10.2478/bsrj-2023-0002.
- [15] M. Shutaywi and N. N. Kachouie, "Silhouette analysis for performance evaluation in machine learning with applications to clustering," *Entropy*, vol. 23, no. 6, p. 759, 2021, doi: 10.3390/e23060759.